Andrey Andreev,
Nikolay Kirov
(Institute of Mathematics and Informatics
Bulgarian Academy of Sciences)

WORD IMAGE MATCHING IN BULGARIAN HISTORICAL DOCUMENTS[1]

Abstract: An approach to word image matching based on Hausdorff distance is examined for bad quality typewritten, printed or handwritten Bulgarian documents. A detailed computer experiments were carried out using 49 pages typewritten text, 13 pages printed text and 2 pages of a manuscript. The results of several methods are compared including previously reported methods in the literature.

Keywords: document text image, bitmap file, word matching, Hausdorff distance

1. Introduction

The Hausdorff distance used in the paper differs slightly from ones used by other authors and after the computer experiments the conclusion from the results is that our method outperforms them despite its simplicity.

Let A and B denote bounded sets on the plane and a and b be points on the plane with coordinates

$$a = (a_1, a_2), \quad b = (b_1, b_2).$$

The Hausdorff distance (HD) between two bounded sets A and B is defined in [4] for the purposes of approximation of discontinues functions as

$$r(A,B) = \max\{h(A,B), h(B,A)\}, \tag{1}$$

where

$$h(A,B) = \max_{a \in A} \min_{b \in B} p(a,b), \tag{2}$$

$$p(a,b) = \max \{ |a_1 - b_1|, |a_2 - b_2| \}. \tag{3}$$

In 1994 Dubuisson and Jain [1] examined 24 distance measures of Hausdorff type for determination to what extend two point sets on the plane A and B differ. In case when the sets A and B consist of N_A and N_B points along with (3) changed to Euclidean distance they use

[1] This research has been supported by a Marie Curie Fellowship of the European Community program “Knowledge Transfer for Digitalization of Cultural and Scientific Heritage in Bulgaria” under contract number MTKD-CT-2004-509754. The work has been done while the authors were at the National Center of Scientific Research “Demokritos”, Athens, Greece.

(4) $$h(A,B) = 1/N_A \sum_{a \epsilon A} \min\{p(a,b): b \epsilon B\}$$

and claim that among all 24 "distances" examined by them, this "distance" called by them "Modified Hausdorff Distance" (MHD) suites in best way the problem for object matching. Similar approach called "Weighted Hausdorff Distance" (WHD) is used in [2] for finding word image matching method in English and Chinese document images. We propose to simplify (4) using slightly changing the definition of h(A,B)

(5) $$h(A,B) = \sum_{a \epsilon A} \min\{p(a,b): b \epsilon B\}$$

and call the distance (1), (2), (5) Sum (or Simple) Hausdorff Distance (SHD).

2. Distances used in computer experiments

Let us define:

- "word image" – a rectangular image which pixels have values 0 (white) or 1 (black);
- "word" – a subset of word image with pixel values equal to 1.

The following distances will be tested numerically for estimation of similarity between two words A and B:

1. $L_1(A,B) = \sum_{a \epsilon (A \backslash B) U (B \backslash A)} 1$;
2. HD(A,B) = r(A,B), where r(A,B) is defined by (1), (2) and (3);
3. $HD_1(A,B)$ = r(A,B), where r(A,B) is defined by (1), (5) and using p(A,B) = 0 if a=b, else p(A,B) = 1;
4. MHD(A, B) = r(A,B), where r(A,B) is defined by (1), (3), (4);
5. SHD(A,B)= r(A,B), where r(A,B) is defined by (1), (3), (5).

Before using a given distance for estimation the difference between two images they must be adjusted with respect to either their geometric centers or to their mass centers. For example if SHD distance is applied combined with geometric center adjustment of images we denote this by SHD^{gc} otherwise we write SHD^{mc}. Measuring the effectiveness of the distances (or methods connected with them) usually is given by standard estimations **Recall** and **Precision** [3]. Briefly, let us look for a word W in a collection of binary text images in which W occurs N times. Let a method produce a sequence of words

(6) $$\{W_i\}_{i=1,2,...}$$

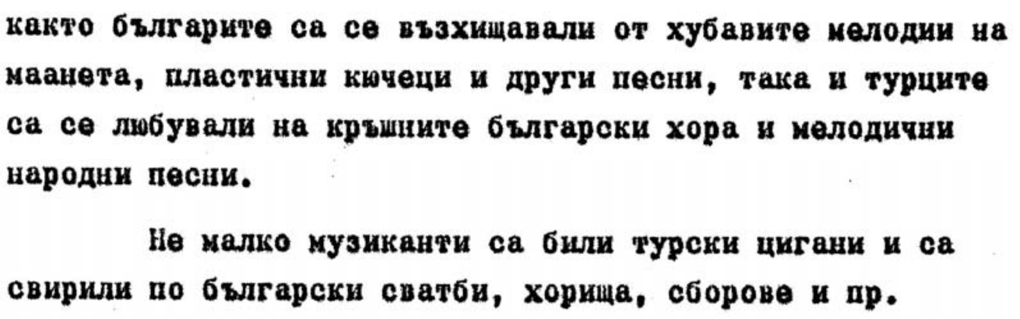
ordered according to a specific criteria. For a given n (n = 1, 2, ...), let n_1 (n_1 is less or equal to n) be the number of words among the first n words in the sequence (6) that coincide with W. Note that n_1 is a function of n. Then we define the following two functions

(7) $$\text{Recall}(n) = n_1 / N \text{ and } \text{Precision}(n) = n_1 / n$$

as functions of n.

3. Experimental results

3.1. Typewritten text. Using the distances defined above we carry out a series of computer word matching experiments. Real Bulgarian documents of typewritten text of 49 pages of bad quality are the material from which a specified word is located and extracted.



както българите са се възхищавали от хубавите мелодии на маанета, пластични кючеци и други песни, така и турците са се любували на кръшните български хора и мелодични народни песни.

Не малко музиканти са били турски цигани и са свирили по български сватби, хорища, сборове и пр.

Word “песни” is the sixth word in the second row in the text above. It occurs 31 times in the whole document of 49 pages.

Series 2: MHD^{gc}			
n	n_1	Recall	Precision
5	5	0.16	1.00
15	14	0.45	0.93
25	21	0.68	0.84
35	25	0.81	0.71
45	27	0.87	0.60
56	30	0.97	0.54
62	31	1.00	0.50

Series 3:HD_1^{gc}			
n	n_1	Recall	Precision
5	5	0.16	1.00
15	14	0.45	0.93
25	20	0.65	0.80
35	24	0.77	0.69
45	25	0.81	0.56
56	28	0.90	0.59
62	29	0.94	0.47
72	30	0.97	0.42
74	31	1.00	0.42

Series 1: SHD^{gc}			
n	n_1	Recall	Precision
5	5	0.16	1.00
15	14	0.45	0.93
25	20	0.64	0.80
35	26	0.84	0.74
45	30	0.97	0.67
56	31	1.00	0.55

Series 4: HD^{gc}			
n	n_1	Recall	Precision
3	3	0.10	1.00
32	26	0.84	0.81
83	31	1.00	0.37

Table 1. Results for “песни”

The results from Table 1 are plotted on Fig.1:

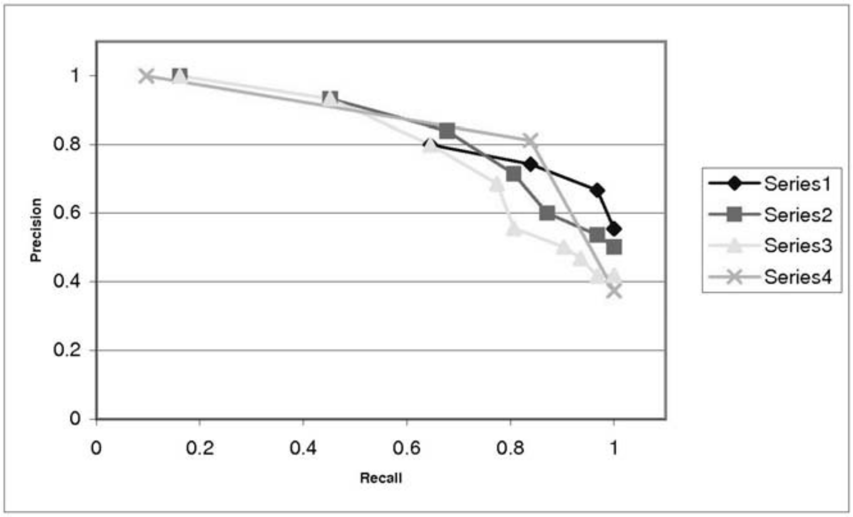
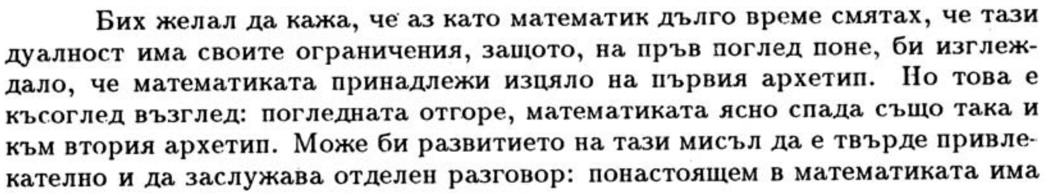


Fig. 1. Word “песни”

3.2. Printed text. We process 13 pages that contain printed text written in proportional font as shown below.



Бих желал да кажа, че аз като математик дълго време смятах, че тази дуалност има своите ограничения, защото, на пръв поглед поне, би изглеждало, че математиката принадлежи изцяло на първия архетип. Но това е късоглед възглед: погледната отгоре, математиката ясно спада също така и към втория архетип. Може би развитието на тази мисъл да е твърде привлекателно и да заслужава отделен разговор: понастоящем в математиката има

The test word “математиката” (mathematics) can be found in the text above as a word before the last one. Our searching template “математик” (mathematician) consists only of the first 9 letters that form it. Thus we can find all words that have the same 9 letters at the beginning. The results of the word recognition process depend on several factors:

- differences between identical words in original document;

Series 1: SHD^{gc}					Series 2:SHD^{gc}			
n	n_1	Recall	Precision		n	n_1	Recall	Precision
9	9	0.26	1.00		18	18	0.53	1.00
16	15	0.44	0.94		25	22	0.65	0.88
22	20	0.59	0.91		32	24	0.71	0.75
29	23	0.68	0.79		47	25	0.74	0.53
37	25	0.74	0.68					
45	28	0.82	0.62					
48	29	0.85	0.60					

Table 2. Results for “математик”

- differences between identical words obtained during the scanning process;
- binarization step (grayscale-b/w conversion and noise removal) can add extra differences;

- segmentation step may extract identical words in a different way.

Our example includes two different binarization steps that give two different recognition results. They are presented in Table 2 and Fig.2 as Series 1 and Series 2.

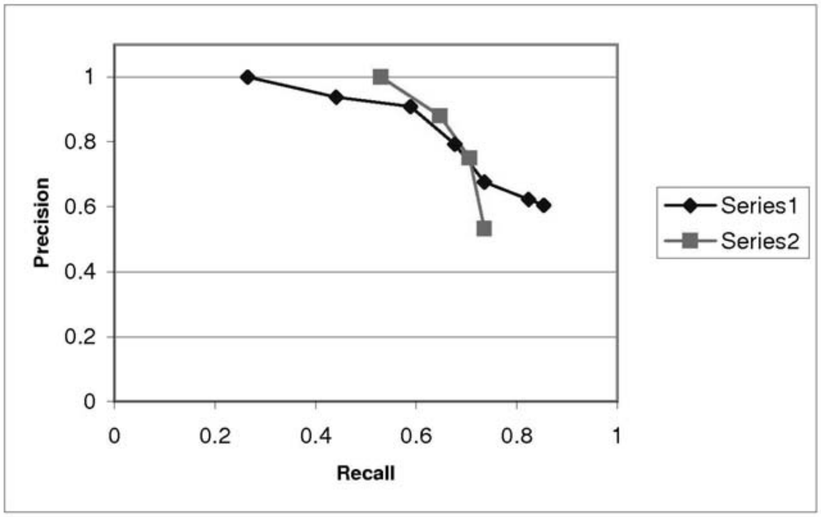
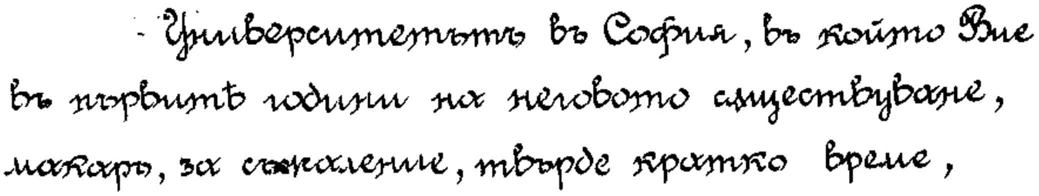


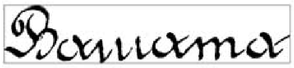
Fig. 2. Word “математик”

3.3. Handwritten text. Two handwritten pages from a document created in 1929 are processed. Part of the document is given below shows the regularity of the text and its quality.

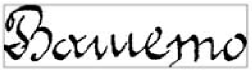


- Университетътъ въ София, въ който Вие
въ първитѣ години на неговото съществуване,
макаръ, за съжаление, твърде кратко време,

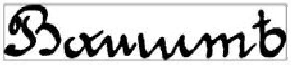
We search for words like the next 4 words that begin with the pattern “Ваш” (“Ваш” means “Yours”).



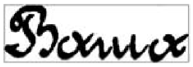
Вашата



Вашето



Вашитѣ



Вашия

The computer results for both methods SHD and HD are summarized in Table 3 and plotted on Fig. 3.

Series 1: SHD^{gc}					Series 2:HD^{gc}			
n	n_1	Recall	Precision		n	n_1	Recall	Precision
2	2	0.22	1.00		5	5	0.56	1.00
4	3	0.33	0.75		8	7	0.78	0.86
7	6	0.66	0.86		10	8	0.89	0.80

10	7	0.78	0.70		27	9	1.00	0.33
13	8	0.89	0.62					
16	9	1.00	0.56					

Table 3. Results for “Ваш”

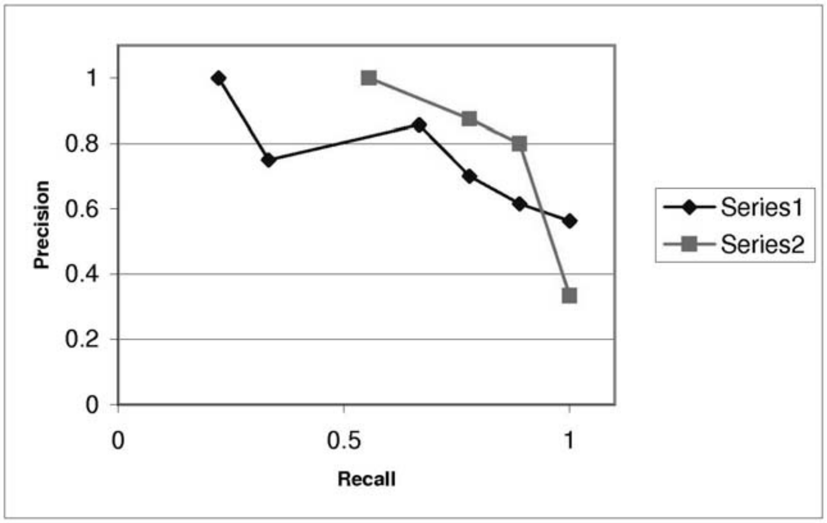


Fig. 3. Word “Ваш”

3. Conclusions

We process bad typewritten Bulgarian text, printed text and a manuscript for word matching using various distances. The results show that:

- the distance SHD^{gc} produces better results than other distances and therefore there is no need to complicate the definition of SHD (like MHD or WHD);
- mass centered adjustment (mc) of word images is inappropriate for the purpose of word matching and we omit computer results obtained in this way;
- L_1^{gc} distance produces the worst results. HD_1^{gc} method which is a sort of a combination of L_1^{gc} and SHD^{gc} behaves better, but evidently falls back to SHD^{gc};
- the measurement done by HD^{gc} distance could be considered as a “discontinuity”. This explains the deterioration of the results produced by HD^{gc} for values of Recall(n) close to 1. For example, computer results for the relatively short word “песни” with occurrence 31 times HD^{gc} are shown in Table 4. Note that in some extend the other methods use practically continuous scale for ordering the spotted words.

HD^{gc} distance	found words - n	correct words – n_1
2	2	2
3	29	23
4	31	5

Table 4. Word “песни” and HD^{gc} distance

References

[1] M.-P. Dubuisson, A. Jain, *A Modified Hausdorff Distance for Object Matching*, In: Proc. 12th Int. Conf. Pattern Recognition, Jerusalem, Israel, 1994, pp. 566–568.

[2] Y. Lue, C. L. Tan, W. Huang, L. Fan, *An Approach to Word Image Matching Based on Weighted Hausdorff Distance*, “6th ICDAR”, 10–13 Sept. 2001, Seattle, USA.

[3] M. Junker, R. Hoch, A. Dengel, *On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy*, Proceedings ICDAR 99, Fifth Intl. Conference on Document Analysis and Recognition, Bangelore, India, 1999.

[4] [4] Bl. Sendov, *Hausdorff Approximations*, Kluwer, Dordreht, 1990.

aandreev@math.bas.bg
nkirov@math.bas.bg